適応制御について (非線型及び線型制御研究会報告集)

AUTHOR(S):

有田, 清三郎; 横田, 敏雄; 坂和, 愛幸

CITATION:

有田, 清三郎 ...[et al]. 適応制御について (非線型及び線型制御研究会報告集). 数理解析研究所講究録 1968, 48: 107-130

ISSUE DATE:

1968-07

URL:

http://hdl.handle.net/2433/107713

数理解析研究所講究録
48巻 1968年 107-130

適応制御について

阪大　基礎工　有田　清三郎
阪大　基礎工　横田　敏雄
阪大　基礎工　坂和　愛幸

本稿では，適応制御への種々のアプローチを紹介し，あわせて，それらの間の関係を述べる。

I　適応制御系の概念とその構成

実際の制御対象の特性は一般に未知であり，可変的である。その不確定さを制御の進行過程で学習させながら希望のふるまいをさせるように制御を行う制御方式を適応制御という。

適応制御系として Fig. 1 のような構成が考えられる。

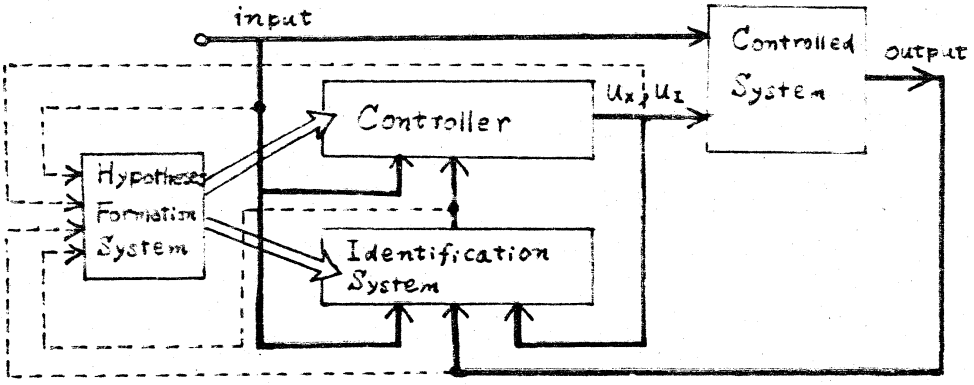

Fig. 1

Bellman [1], 深尾 [2], [3] 等は，従来の「物理的な状態」を表わす変数Xに「知識の状態」を表わす変数Iをつけ加え，

$$X'(t_2) = X'(X(t_1), I(t_1), u_X(t_1), u_I(t_1), \gamma(t_1), t_2)$$
$$I'(t_2) = I'(X(t_1), I(t_1), u_X(t_1), u_I(t_1), \gamma(t_1), t_2) \qquad (1.1)$$

の如き“物理－情報統合系”を考えている。

ここに，$X(t_1)$：時刻 t_1 の物理的状態，$I(t_1)$：時刻 t_1 までの過去の経験の蓄積によって得た知識の状態，$X'(t_2)$：制御が行なわれる時刻 t_2 での物理的状態の予測値，$I'(t_2)$：時刻 t_2 での知識の状態の予測値，$\gamma(t_1)$：時刻 t_1 での未知確率構造をもつ確率変数，$u_X(t_1)$：時刻 t_1 での物理系の制御（制御対象に対して加えられる制御），$u_I(t_1)$：時刻 t_1 での情報系の制御（知識向上のための情報入手に際する実験の選択　観測の選択，情報処理方式の変更等を意味する制御）。

また，評価関数には X, u_X, u_I, I が加わるのが一般的である。従って，適応制御系における adaptive optimal controller は未知のものに対する学習と最適制御の2つの機能を組合せたものでなくてはならない。この考え方は Feldbaum [4] [5] の“dual control”の概念とも一致する。

Ⅱ. 適応制御に対する種々のアプローチ.

制御系の不確定さは大別すると，次の2つになる。

(A) 未知パラメータだけに依存する場合，すなわち未知パラメータが求まれば，系は確定的ないしは確率的にきまってしまう場合．(B) 未知パラメータだけに依存しない場合（たとえばシステムの構造が未知な場合）．

(A)に対するアプローチには (1) Bayes approach (2) Min-max approach の2通りが考えられる。(1)は未知パラメータについての先験的情報が与えられている場合であり，(2)は先験的情報が未知パラメータの属する空間のみの場合である。

(B)に対するアプローチには(3) Stochastic approximation，(4) Reinforcement learning algorithm (5) Potential function method 等が考えられる。

§1. Bayes approach

次のような離散系

$$X_{k+1} = F_k(X_k, u_k, v_k) \qquad k=0,1,\cdots,N-1 \quad (2.1)$$

$$Y_k = G_k(X_k, \xi_k) \qquad k=1,2,\cdots,N \quad (2.2)$$

ここに X_k, u_k は時刻 k における状態及び制御であり，X^k, u^k はそれぞれ系列 $\{X_1, X_2, \cdots, X_k\}$，$\{u_1, u_2, \cdots, u_k\}$ を表わす。v_k は未知パラメータ θ を母数とする確率構造に従う確率変数，Y_k は観測値，ξ_k は雑音である。F_k の関数形は既知とする。評価関数は，制御の系列 u^N と確率変数 v^N の関数として考え

られるから，これを $J(u^N, v^N)$ と書く。たとえば

$$J(u^N, v^N) = \sum_{i=1}^{N} W_i(u_{i-1}, x_i) \qquad (2.3)$$

なるものが考えられる。ここに N は固定されており，W_i は既知関数である。

上のように，不確定さが未知パラメータに内在しており，かつその未知パラメータの先験的分布が与えられている場合には Bellman [1], Feldbaum [4],[5], Aoki [6], Sworder [7] 等の提案した Bayes approach が有効である。

制御 u，ランダム変数 v の確率密度関数をそれぞれ $\xi(u), p(v)$ とし，ξ の集合を Ξ とする。ある $p(v)$ に対して

$$\iint J(u,v)\xi^\circ(u)p(v)\,du\,dv = \inf_{\xi\in\Xi} \iint J(u,v)\xi(u)p(v)\,du\,dv$$

なる ξ° を $p(v)$ に対する Bayes 最適制御という。制御が純粋方略をとる場合，u の集合を U とすると，Bayes 最適制御 u° は次式を満たす。

$$\int J(u^\circ, v)p(v)\,dv = \inf_{u\in U} \int J(u,v)p(v)\,dv \qquad (2.4)$$

統計的決定理論での $p(v)$ は既知としているのに対して，適応制御における Bayes approach では $p(v)$ が未知なので，$p(v)$ をそれまでの観測データによって学習し，変更させていく事に大きな違いがある。その学習は次のようにして行なう。

与えられた先験的確率密度 $p_0(v)$ から出発して，Bayes の定理によって，観測 Y_1 の後の事後確率密度

$$p(v|Y_1) = P(Y_1|v)\,p_0(v) \Big/ \int p(Y_1|v)\,p_0(v)\,dv \qquad (2.5)$$

が得られる。このような手順を逐次くりかえすことによって観測の系列 $Y^n=\{Y_1,Y_2,\cdots,Y_n\}$ が得られた後の事後確率密度 $P(v\mid Y^n)\triangleq P_n(v)$ は次のように与えられる。

$$P_n(v)=P(Y_n\mid v)P_{n-1}(v)\Big/\int P(Y_n\mid v)P_{n-1}(v)dv \qquad (2.6)$$

$$(n=1,2,\cdots\cdots)$$

このようにして得られた $P_n(v)$ と制御の確率密度関数 $\xi_n(u)$ に対して，$\iint J(u,v)\xi_n(u)P_n(v)dudv\triangleq E_{\xi_n P_n}[J(u,v)]$ と定義すると，次の期待値

$$E_{\xi_1 P_1}\cdots E_{\xi_N P_N}[J(u^N,v^N)]=E_{\xi_1 P_1}\cdots E_{\xi_N P_N}\Big[\sum_{i=1}^{N}W_i\Big]$$

$$=E_{\xi_1 P_1}\Big(W_1+E_{\xi_2 P_2}\big(W_2+E_{\xi_3 P_3}(W_3+\cdots+E_{\xi_N P_N}(W_N))\cdots\big)\Big) \qquad (2.7)$$

を最小にするような $\xi_1,\xi_2,\cdots,\xi_n$ が Bayes 最適制御である。このような Bayes 最適制御は一般にランダム方略になるのであるが，Feldbaum [4] はこれが純粋方略になる事を示している。

註）(1) 他のアプローチが一段毎での最適制御しか得られないのにたし，Bayes approach では多段過程の最適制御を得ることができる。この事は他のアプローチにくらべてきわめて有効である。(2) Bayes approach の大きな特徴は，先験的確率密度を与えなければならないことである。しかしながら，この先験的確率密度は一般には未知であり，これを正しく設定することはむつかしい。この欠点を補うため，深尾 [2] は次のよう

なアプローチを提案している。すなわち，ランダム変数 v についての先験的確率密度として，いくつかの可能な確率密度関数 $\eta_0^1(v), \eta_0^2(v), \cdots, \eta_0^S(v)$ をとって，各々の密度関数へとる確率を $p_1, p_2, \cdots, p_S$ とする。(p_i は何らかの方法できめる。また p_i を学習させてもよい。) ここに，$\sum_{i=1}^{S} p_i = 1$, $p_i \geq 0$ $(i=1,2,\cdots,S)$. このとき，新しい評価関数

$$\sum_{i=1}^{S} p_i E^i_{\eta_1} \cdots E^i_{\eta_N} \left[J(U^N, V^N) \right] \qquad (2.8)$$

を最小にする最適制御を求める。ここに $E^i_{\eta_n}$ は観測の系列 $Y^n = \{Y_1, Y_2, \cdots, Y_n\}$ を得た後の確率密度関数 η_0^i の事後確率密度 η_n^i による平均操作を表わす。η_n^i は (2.6) と同様にして計算することができる。

§2. Min-max approach

制御 u，ランダム変数 v の確率密度関数をそれぞれ $\xi(u)$, $\eta(v)$ とし，ξ, η の集合をそれぞれ Ξ, H とする。また，

$$J(\xi, \eta) \triangleq \iint J(u,v)\xi(u)\eta(v)\,du\,dv \qquad (2.9)$$

と定義するとき，

$$\sup_{\eta \in H} J(\xi^\circ, \eta) = \inf_{\xi \in \Xi} \sup_{\eta \in H} J(\xi, \eta) \qquad (2.10)$$

なる ξ° を min-max policy という。

制御 u，ランダム変数 v がともに純粋方略をとる場合には，(2.10) において $\xi(u)=\delta(u)$，$\eta(v)=\delta(u)$（ただし $\delta(\cdot)$ はデルタ関数を表わす。）とおくことにより，純粋方略における min-max policy u° は，次のように定義できる。

$$\inf_{u\in U}\sup_{v\in V} J(u,v)=\sup_{v\in V} J(u^\circ,v) \qquad (2.11)$$

ここに，U，V はそれぞれ純粋方略 u，v の集合である。

Bayes policy と min-max policy との関係は，次の諸定理で示される。[7],[8]

[定理 1]
$$\inf_{\xi\in\Xi}\sup_{\eta\in H} \bar{J}(\xi,\eta)=\sup_{\eta\in H}\inf_{\xi\in\Xi} \bar{J}(\xi,\eta) \qquad (2.12)$$

が成立し，かつ least favorable distribution η° すなわち
$$\inf_{\xi\in\Xi} \bar{J}(\xi,\eta^\circ)=\sup_{\eta\in H}\inf_{\xi\in\Xi} \bar{J}(\xi,\eta) \qquad (2.13)$$

なる η° が存在するならば，min-max policy ξ° は η° に対する Bayes policy である。

[定理 2] ξ° が η° に対する Bayes policy でかつ，すべての $v\in V$ に対して
$$\int J(u,v)\xi^\circ(u)\,du \le \bar{J}(\xi^\circ,\eta^\circ) \qquad (2.14)$$
が成立するとき，

(1) $$\inf_{\xi\in\Xi}\sup_{\eta\in H} \bar{J}(\xi,\eta)=\sup_{\eta\in H}\inf_{\xi\in\Xi} \bar{J}(\xi,\eta)$$

(2) ξ^0 は min-max policy である。

(3) η^0 は least favorable である。

[定理3] ξ_n が η_n に対する Bayes policy で，$n\to\infty$ のとき $r(\xi_n,\eta_n)\to C$，すべての v に対して $\int J(u,v)\xi^0(u)du \le C$ ならば，上の定理2の (1),(2) が成りたつ。（証明略）

任意の $\varepsilon>0$ に対して

$$r(\xi^0,\eta) \le \inf_{\xi\in\Xi} r(\xi,\eta)+\varepsilon$$

なる η が存在するとき，ξ^0 を extended Bayes policy という。

また，すべての $v\in V$ に対して

$$\int J(u,v)\xi^0(u)du = C \quad (C\text{はある定数})$$

なる ξ^0 が存在するとき，ξ^0 を equalizer policy という。

[定理4] ξ^0 が equalizer でかつ extended Bayes ならば，ξ^0 は min-max policy である。

註) (1) min-max approach は Bayes approach のような先験的情報を必要としない。(2) この方法は，最悪の事態を予測した場合の安全策であるか，不確定さは意識的な敵対者ではないから，この方法は消極的であると見做される。(3) equalizer policy は，v に対して独立であるから，ランダム変数に対して insensitive な最適制御と解される。

§3 Stochastic approximation

定常のランダムベクトル X と N 次元パラメータベクトル $\mathbf{c}$ の関数 $Q(X|\mathbf{c})$ を考える。X の密度関数 $p(x)$ が与えられているとして，Q の X についての期待値

$$I(\mathbf{c}) \triangleq \int_{\Omega} Q(x|\mathbf{c})\,p(x)\,dx \triangleq E_x\{Q(x|\mathbf{c})\} \qquad (2.17)$$

の極値を求める問題を考えてみよう。

$I(\mathbf{c})$ が $\mathbf{c}$ について微分可能ならば，(2.17) の極値を与える $\mathbf{c}^*$ は

$$\nabla I(\mathbf{c}) = \nabla_{\mathbf{c}} E_x\{Q(x|\mathbf{c})\} = 0 \qquad (2.18)$$

を満足する。$p(x)$ が既知ならば，勾配法によって (2.18) を満足する最適ベクトル $\mathbf{c}^*$ を求めることができるが，$p(x)$ は一般に未知であるから，Q が $\mathbf{c}$ について微分可能のとき，$\nabla_{\mathbf{c}} E_x\{Q(x|\mathbf{c})\}$ の代りに $\nabla_{\mathbf{c}} Q(x|\mathbf{c})$ についての勾配法を適用して $\mathbf{c}^*$ を逐次的にきめる方法が *Stochastic approximation* である。

この方法は，1951 年 Robbins-Monro [9] によって回帰方程式の解を求める問題に対して開発され，1952 年 Kiefer-Wolfowitz [10] によって回帰関数の極値を求める問題に拡張された。その後，収束条件について検討がなされ，Dvoretzky [11] によって要約された。

$\mathbf{c}^*$ を求めるアルゴリズムは，$\sum_{n=1}^{\infty} \gamma_n = \infty$，$\sum_{n=1}^{\infty} \gamma_n^2 < \infty$ なる条件を満足する実数 γ_n を用いて，次のように与えられる。

Q が $\mathbf{c}$ について微分可能なとき

$$\mathbf{c}_n = \mathbf{c}_{n-1} + \gamma_n \nabla_{\mathbf{c}} Q(X_n | \mathbf{c}_{n-1}) \qquad (2.19)$$

Q が $\mathbf{c}$ について微分可能でないときも同様なアルゴリズムが得られる。[12] (2.19) アルゴリズムを *regular type* (*Robbins-Monro processure*), Q が微分可能でないときのアルゴリズムを *irregular type* (*Kiefer-Wolfowitz processure*) という。このようなアルゴリズムは，適当な条件のもとで，真の値 $\mathbf{c}^*$ に確率1の収束及び平均収束することが知られている。[11]. (*Appendix I* 参照)

次に，この方法を適応制御問題に応用しよう。

$$X_n = f(X_{n-1}, U_{n-1}) \qquad (2.20)$$

なる制御対象に対して，

$$U_n = g(X_n) \qquad (2.21)$$

なる形の制御をほどこすとき，評価関数

$$I_1 = E[F_1(X_n^\circ - X_n)] \qquad (2.22)$$

を最小にするような制御を求める問題を考えよう。

ここに $f(\cdot,\cdot)$ は未知関数，$F_1(\cdot)$ は既知な凸関数，E は期待値，X_n° $(n=1,2,\cdots)$ は与えられた希望値を表わす。

Tsypkin[12],[13]はこの問題に対して *Stochastic approximation* を用いて，次のような解法を示している。まず，未知関数 $f(\cdot,\cdot)$, $g(\cdot)$ を次のような一次独立な既知関数 φ_i, ψ_j の一次

結合で近似する。すなわち

$$f(x,u) \cong \hat{f}(x,u) = \sum_{i=1}^{p} c_i \phi_i(x,u) = \mathbf{c}'\boldsymbol{\Phi}(x,u) \quad (2.23)$$

$$g(x) \cong \hat{g}(x) = \sum_{j=1}^{q} b_j \psi_j(x) = \mathbf{b}'\boldsymbol{\Psi}(x) \quad (2.24)$$

ここに，b_j, c_i は未知パラメータであり，$\mathbf{c} = (c_1, \cdots, c_p)$ $\mathbf{b} = (b_1, \cdots, b_q)$，$\boldsymbol{\Phi} = (\phi_1, \cdots, \phi_p)$，$\boldsymbol{\Psi} = (\psi_1, \cdots, \psi_q)$.

この問題では，制御対象たる f が未知であるから，f の identification と最適制御とを考えねばならない。制御対象を学習するために，新たに系の identification に対する誤差評価関数を導入し，それを最小にすることを考える。すなわち F_2 を適当な凸関数とし

$$I_2(\mathbf{c}) = E\{F_2[x_n - \mathbf{c}'\boldsymbol{\Phi}(x_{n-1}, u_{n-1})]\} \quad (2.25)$$

とおく。(2.25)を最小にするアルゴリズムは

$$\mathbf{c}_n = \mathbf{c}_{n-1} + \gamma_n' \nabla_{\mathbf{c}} F_2[x_n - \mathbf{c}_{n-1}'\boldsymbol{\Phi}(x_{n-1}, u_{n-1})] \quad (2.26)$$

従って，評価関数 I_1 は

$$I_1(\mathbf{b}, \mathbf{c}) = E\{F_1[x_n^\circ - \mathbf{c}'\boldsymbol{\Phi}(x_{n-1}, \boldsymbol{\Psi}'(x_{n-1})\mathbf{b})]\} \quad (2.27)$$

となる。この $I_1(\mathbf{b}, \mathbf{c})$ を $\mathbf{b}$ について最小にするようなアルゴリズムは

$$\mathbf{b}_n = \mathbf{b}_{n-1} + \gamma_n \nabla_{\mathbf{b}} F_1[x_n^\circ - \mathbf{c}_{n-1}'\boldsymbol{\Phi}(x_{n-1}, \boldsymbol{\Psi}'(x_{n-1})\mathbf{b}_{n-1})] \quad (2.28)$$

$\mathbf{b}_n$, $\mathbf{c}_n$ のアルゴリズム (2.26), (2.28) を互いにくりかえし計算することによって，最適な $\mathbf{b}^*$, $\mathbf{c}^*$ を見つけることができる。

Tsypkin は Fig.2 の構成で b_n, c_n を逐次求めることを提案している。

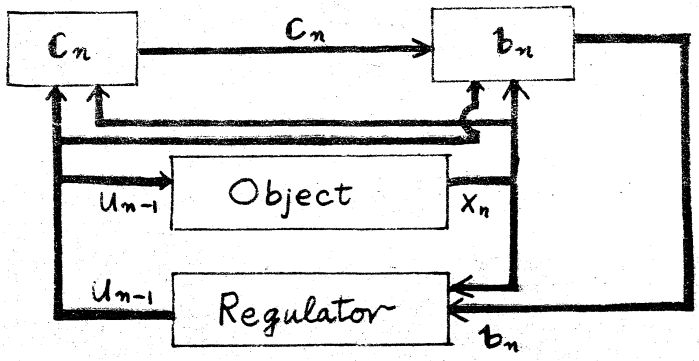

Fig.2 Tsypkin の adaptive control system

註) (1) この方法は，制御対象の関数形が未知な場合に，汎関数を一次独立な関数の一次結合で近似することによって，汎関数の最小問題を最適係数ベクトルを求める問題に還元している。またこのアルゴリズムは簡単であり，計算機にかかりやすい。以上の点から，この方法は適応制御に有効である。(2). この方法は一段毎の評価をとっていくことにより，多段の最適制御を行なおうとするものである。これは多段制御過程での sub-optimal しか得られないことを意味する。(3). この方法では，$n \to \infty$ のとき，真の値に確率1の収束をするからこの方法の実際面の適用においては，多数回のくりかえし計算が必要である。

§4. Reinforcement learning algorithm [14]

Reinforcement learningは stochastic automaton の learning model の一つである。その基本的原理は，得られた状態確率にもとづく評価量のうち，最大(又は最小)なものの状態確率を補強し，他の状態確率を減ずるという帰納的かつ直観的な学習である。

Reinforcement learning algorithmは学習の仕方によって linear と non-linear とがあるが，ここでは linear なものを紹介し，それを適応制御に応用する。

<Linear reinforcement learning algorithm>

時刻 n における状態 $g(n)$ が g_i である確率を $P_n(i)$ とする。すなわち

$$P_n(i) = P\{g(n) = g_i\} \tag{2.29}$$

linear reinforcement learning algorithm の大きな特徴は $P_n(i)$ と $P_{n-1}(i)$ がすべての i に対して線形の関係にあることである。すなわち

$$P_n(i) = \alpha_n P_{n-1}(i) + (1-\alpha_n)\lambda_n(i) \quad (i=1,2,\cdots S) \tag{2.30}$$

ここに $$0 \le P_n(i) \le 1, \quad \sum_{i=1}^{S} P_n(i) = 1 \tag{2.31}$$

$$0 \le \lambda_n(i) \le 1, \quad \sum_{i=1}^{S} \lambda_n(i) = 1 \tag{2.32}$$

また α_n は $0 < \alpha_n < 1$ を満足する実数，$\lambda_n(i)$ は時刻 n での観測にもとづく補強値である。(2.30)から，$P_n(i)$ が，$P_{n-1}(i)$ と情報を得た後の補強値 $\lambda_n(i)$ との加重平均になっていることが

わかる。

〈適応制御問題への応用〉

まず，Fu [14] の linear reinforcement learning algorithm の learning control への応用を示そう。

時刻 n における状態，制御，観測をそれぞれ $q(n) \in Q$，$u(n) \in U$, $y(n) \in Y$ とする。

いま，
$$q(n+1) = F[q(n), u(n)] \tag{2.33}$$
なる制御対象に対して
$$u(n) \in U = \{u_1, u_2, \cdots, u_m\} \tag{2.34}$$
なる制御をほどこす時，評価関数
$$E\{J(q(n+1), y(n), u(n)) \mid y(n), u(n)\} \tag{2.35}$$
を最小にするような $u(n)$ を求める問題を考える。ここに，F は未知の関数，J は既知関数，E は期待値を表わす。

上の問題で，実際には J の確率分布は未知であるから，$E\{J \mid y, u_i\}$ を
$$P\left[\lim_{n_i \to \infty} \hat{E}_{n_i}\{J \mid y, u_i\} = E\{J \mid y, u_i\}\right] = 1 \tag{2.36}$$
なる性質をもった逐次推定値 $\hat{E}_{n_i}\{J \mid y, u_i\}$ から on-line で推測する。ここに n_i は u_i をほどこした回数である。$E\{J \mid y, u_i\}$ の逐次推定値 $\hat{E}_{n_i}\{J \mid y, u_i\}$ は，たとえば次のようにとれる。
$$\hat{E}_{n_i}\{J \mid y, u_i\} = \frac{1}{n_i} \sum_{m=1}^{n} \delta_m J(q(m+1), y(m), u(m))$$
ここに $\sum_{m=1}^{n} \delta_m = n_i$，$\delta_m = \begin{cases} 1 & (u(m) = u_i \text{ のとき}) \\ 0 & (u(m) \neq u_i \text{ のとき}) \end{cases}$

時刻 n における制御 u_i の確率を次のように定義する。

$$p_n(i) \triangleq P\{u_i \mid y(n)\} \quad (i=1,2,\cdots,m) \qquad (2.37)$$

いま，y が n 回観測され，制御 $u_1,\cdots,u_i,\cdots,u_m$ がそれぞれ n_1 回，$\cdots$，n_i 回，$\cdots$，n_m 回 $\left(n=\sum_{j=1}^{m} n_j\right)$ ほどこされた後，

$$\hat{E}_{n_k}[J \mid y, u_k] = \min\left\{\hat{E}_{n_j}[J \mid y, u_j] : j=1,\cdots,m\right\} \qquad (2.38)$$

とする。このとき時刻 n における制御 u_i の確率 $p_n(i)$ を *linear reinforcement learning algorithm* で次のように推定しよう。

$$p_n(i) = \alpha_n p_{n-1}(i) + (1-\alpha_n)\lambda_n(y, u_i) \qquad (2.39)$$

ここに $\lambda_n(y,u_i) = \begin{cases} 1 & (i=k \text{ のとき}) \\ 0 & (i \neq k \text{ のとき}) \end{cases}$ $(n=1,2,\cdots\cdots)$ $\qquad (2.40)$

従って，$i=k$ のとき

$$p_n(i) = \alpha_n p_{n-1}(i) + (1-\alpha_n) > p_{n-1}(i) \qquad (2.41)$$

$i \neq k$ のとき

$$p_n(i) = \alpha_n p_{n-1}(i) < p_{n-1}(i) \qquad (2.42)$$

となり，(2.38) に対応する制御 u_k の確率を補強し，他の制御の確率を減じていることがわかる。

この *learning control* は，(2.39) の $p_n(i)$ が，($E\{J \mid y, u_i\}$ が既知であれば) $\sum_{i=1}^{m} p(i) E\{J \mid y, u_i\}$ を最小にする事を意味している。

次に，linear reinforcement learning algorithm を使って適応制御系の一例を考えてみよう。

$Q_A \equiv \{g_0\}$ とし，固定された状態 g_0 から状態の集合 $Q_B \equiv \{g_1, \cdots, g_S\}$ への状態遷移確率ベクトルを T とすると，その成分は g_0 から g_i $(i=1,2,\cdots,S)$ への遷移確率 $p(i)$ である。

$$g_i \in Q_B = T Q_A = \begin{pmatrix} p(1) \\ \vdots \\ p(S) \end{pmatrix} g_0 \qquad (2\cdot43)$$

に対して $$u \in U \equiv \{u;\ 0 \le u \le 1\} \qquad (2\cdot44)$$

なる制御をほどこし，評価関数

$$\sum_{i=1}^{S} p(i) J(g_i, u) \qquad (2\cdot45)$$

を最小にするような最適制御を求める問題を考える。ここに $p(i)(i=1,2,\cdots,S)$ は未知なる遷移確率，J は既知関数とする。

$p(i)$ は未知だから，時刻 n における $p(i)$ の推定値を $\hat{p}_n(i)$ とすると，$\hat{p}_n(i)$ にもとづいて (2·45) を最小にするような最適制御を求めればよい。ここで $\hat{p}_n(i)$ を linear reinforcement learning algorithm で学習することにする。すなわち

$$\hat{p}_n(i) = \alpha_{n-1}\, \hat{p}_{n-1}(i) + (1-\alpha_{n-1})\lambda_{n-1}(i) \qquad (2\cdot46)$$
$$(n=1,2,\cdots\cdots)$$

ここに，$\hat{p}_0(i)$ はあらかじめきめておく。たとえば $\hat{p}_0(i)=\frac{1}{S}$ $(i=1,2,\cdots,S)$。また，$\lambda_n(i) = n_i/n$ $\qquad (2\cdot47)$

n_iはn回の観測のうちg_iのでた回数であり，$n=\sum_{i=1}^{r} n_i$, $n_i \geq 0$.
また，α_nは $0<\alpha_n<1$， $\prod_{n=1}^{\infty}\alpha_n = 0$， $\sum_{n=1}^{\infty}(1-\alpha_n)^2 < \infty$ (2·48)
を満足するものとする。 (2·55)の$\hat{p}_n(i)$が真の値に確率1の収束をすることは，Ⅲ－§2で示す。

この適応制御系は，$p(i)$の推定と最適制御が分離されているか，$p(i)$が制御uの影響をうけるならば

$$\hat{p}_n(i) = \alpha_{n-1}\hat{p}_{n-1}(i) + (1-\alpha_{n-1})\lambda_n(g_i, u)$$

の如くλ_nがuにも関係してくるような学習を行なえば良い。

註) この方法は実際的であるか，Fu〔14〕の learning control では一段毎での最適制御しか得られない。

Ⅲ 種々のアプローチと Stochastic approximation との関係

§1. Bayes learning と Stoch. approximation との関係

Bayes learning と Stoch. approximation との関係について，Fu and Chein〔15〕は，ある種の Bayes learning が実は Stoch. approximation の一つであること及びその収束性を示している。

ここでは最もわかりやすい，平均値ベクトルM，共分散行列Kの正規分布$N(M.K)$で，Mが未知の場合の Bayes learning と Stoch. approximation の関係を示す。

まず未知母数Mの先験的分布を$N(M_0, \Phi_0)$とすると観測値$X_1, X_2, \ldots, X_n$を得た後の事後分布もやはり正規分布となり，そ

の平均値ベクトル及び共分散行列をそれぞれ M_n, Φ_n とすれば,

$$M_n = \left(\frac{1}{n}K\right)\left(\Phi_0 + \frac{1}{n}K\right)^{-1} M_0 + \Phi_0\left(\Phi_0 + \frac{1}{n}K\right)^{-1} \langle X \rangle \quad (3.1)$$

$$\Phi_n = \left(\frac{1}{n}K\right)\left(\Phi_0 + \frac{1}{n}K\right)^{-1} \Phi_0 \quad (3.2)$$

となる。 ここに $\langle X \rangle = \frac{1}{n}\sum_{i=1}^{n} X_i$ (3.3)

(3.1) から, M_n が, M_0 と sample information $\langle X \rangle$ との加重平均となっていることがわかる。

$$\Phi_0 = \alpha^{-1} K, \quad \alpha > 0 \quad (3.4)$$

とおける。このとき, (3.1), (3.2)は

$$M_n = \frac{\alpha}{n+\alpha} M_0 + \frac{n}{n+\alpha} \langle X \rangle \quad (3.5)$$

$$\Phi_n = \frac{1}{n+\alpha} K \quad (3.6)$$

となる。 $n \to \infty$ のとき, $M_n \to \langle X \rangle$, $\Phi_n \to 0$ となり、M_nは正規分布の真の平均値ベクトルに収束する。

以上を Stoch. approximation の言葉で書きかえてみよう。

(3.5) は M_n, M_{n-1} の漸化式

$$M_n = K(\Phi_{n-1} + K)^{-1} M_{n-1} + \Phi_{n-1}(\Phi_{n-1} + K)^{-1} X_n \quad (3.7)$$

に書きかえることができる。(3.7)を更に変形して

$$M_n = M_{n-1} + \Phi_{n-1}(\Phi_{n-1} + K)^{-1}(X_n - M_{n-1}) \quad (3.8)$$

これは, まさしく Stoch. approximation の一つのアルゴリズムである。 (3.8)の収束項 $\Phi_{n-1}(\Phi_{n-1} + K)^{-1}$ は (3.6) より

$$\Phi_{n-1}(\Phi_{n-1} + K)^{-1} = \frac{1}{n+\alpha} I \quad (I\text{は単位行列}) \quad (3.9)$$

よって, (3.8) は $M_n = M_{n-1} + \frac{1}{n+\alpha}(X_n - M_{n-1}) = M_{n-1} + \gamma_n(X_n - M_{n-1})$ (3.10)

$$\gamma_n = \frac{1}{n+\alpha} \tag{3.11}$$

次に (3·10), (3·11) のアルゴリズムは Dvoretzky の条件 (Appendix I 参照) を満足することが示されるから，M_n は真のベクトル M に確率1の収束及び平均収束する。すなわち

$$P\left(\lim_{n\to\infty} M_n = M\right) = 1 \tag{3.12}$$

$$\lim_{n\to\infty} E\{\|M_n - M\|\} = 0 \tag{3.13}$$

である。以下にこれを示そう。

$$\gamma_n > 0, \quad \lim_{n\to\infty} \gamma_n = 0 \tag{3.14}$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty, \qquad \sum_{n=1}^{\infty} \gamma_n^2 < \infty \tag{3.15}$$

は明らかである。いま $X_n = M + H_n$ とし，H_n は $E\{\lambda_n^i\} = 0$, $E\{(\lambda_n^i)^2\} < \infty$ $(i = 1, 2, \cdots, k)$ なる k 個の成分からなるランダムベクトル $(\lambda_n^1, \lambda_n^2, \cdots, \lambda_n^k)$ とする。このとき (3·10) は

$$M_n = M_{n-1} + \gamma_n (M + H_n - M_{n-1}) = (1-\gamma_n) M_{n-1} + \gamma_n M + \gamma_n H_n \tag{3.16}$$

とかける。さらに $T_n(M_1, M_2, \cdots, M_{n-1}) = (1-\gamma_n) M_{n-1} + \gamma_n M$ とおく。このとき

(D·1)
$$\|T_n(M_1, \cdots, M_{n-1}) - M\| = \|(1-\gamma_n) M_{n-1} + \gamma_n M - M\|$$
$$= (1-\gamma_n)\|M_{n-1} - M\| = F_n \|M_{n-1} - M\| \tag{3.17}$$

である。ここに
$$F_n = 1 - \gamma_n = 1 - \frac{1}{n+\alpha} = \frac{n+\alpha-1}{n+\alpha} \tag{3.18}$$

まず，$F_n > 0$ であることは明らかである。また

(D·2)
$$\prod_{n=1}^{\infty} F_n = \lim_{k\to\infty} \prod_{n=1}^{k} F_n = \lim_{k\to\infty} \frac{\alpha}{k+\alpha} = 0 \tag{3.19}$$

このとき，(D·3)
$$M_n = T_n(M_1, M_2, \cdots, M_{n-1}) + \gamma_n H_n \tag{3.20}$$

$\|M_0\|<\infty$ と仮定してよいし，$E\{\|H_n\|^2\}=\sum_{i=1}^{k_n} E\{(\lambda_n^i)^2\}\le B<\infty$ であるから (D·4)

$$E\{\|M_0\|^2\}+\sum_{n=1}^{\infty}E\{\|\gamma_n H_n\|^2\}$$
$$=E\{\|M_0\|^2\}+\sum_{n=1}^{\infty}\gamma_n^2 E\{\|H_n\|^2\}\le E\{\|M_0\|^2\}+B\sum_{n=1}^{\infty}\gamma_n^2<\infty \quad (3.21)$$

また，(D·5) すべての measurable function $\varphi(M_1,M_2,\cdots,M_n)$ に対して

$$E\{\|\varphi(M_1,\cdots,M_n)+H_n\|^2\}$$
$$\le E\{\|\varphi(M_1,\cdots,M_n)\|^2\}+E\{\|H_n\|^2\} \quad (3.22)$$

が成立する。以上より Dvoretzky の条件 (D·1)～(D·5) が満足されることが示された。

§2 Linear reinforcement learning algorithm と Stoch. approximation との関係

ここでは，linear reinforcement learning algorithm もまた Stoch. approximation のアルゴリズムの一つであることおよびその収束性を示す。【14】

linear reinforcement learning algorithm は，次式で表わされる。

$$p_{n+1}(i)=\alpha_n p_n(i)+(1-\alpha_n)\lambda_n(i)$$
$$=p_n(i)+(1-\alpha_n)(\lambda_n(i)-p_n(i)) \quad (3.23)$$

これは Stoch. approximation のアルゴリズムの一つである。次に (3.23) において，

$$\lambda_n(i)=n_i/n \quad (3.24)$$

とおき，

$$0<\alpha_n<1,\quad \prod_{n=1}^{\infty}\alpha_n=0,\quad \sum_{n=1}^{\infty}(1-\alpha_n)^2<\infty \quad (3.25)$$

と仮定するとき，(3.23) のアルゴリズムは Dvoretzky の条件を

満足し，$p_n(i)$ は真の値 $p(i)$ に確率1の収束をする。以下これを示す。(3.23) は

$$p_{n+1}(i)-p(i)=\alpha_n[p_n(i)-p(i)]+(1-\alpha_n)\eta_n(i) \quad (3.27)$$

と書きあらわすことができる。ここに

$$\eta_n(i)=\lambda_n(i)-p(i)$$
$$E[\eta_n(i)]=0, \quad E[\{\eta_n(i)\}^2]<1 \quad (3.28)$$

(3.27) より $p_{n+1}(i)=p_n(i)+(1-\alpha_n)(\eta_n(i)+p(i)-p_n(i))$

$$=\alpha_n p_n(i)+(1-\alpha_n)p(i)+(1-\alpha_n)\eta_n(i)$$

いま，$T_n(p_1(i),\cdots,p_n(i))=\alpha_n p_n(i)+(1-\alpha_n)p(i)$ (3.29)

とおけば，$[T_n(p_1(i),\cdots,p_n(i))-p(i)]=\alpha_n[p_n(i)-p(i)]$ (3.30)

$F_n=\alpha_n$ とおけば，前節と同様にして Dvoretzky の条件が満足され，従って，$P\{\lim_{n\to\infty} p_n(i)=p(i)\}=1$ (3.31)

が示される。

Ⅳ 結論

(1) 現在考えられている適応制御系では，知識の状態を，確率変数の確率構造を学習するという事で把えているが，適応制御で最も重要な事は，情報の定量化を行なう事である。

(2) 実際面の適用としては *Stochastic approximation*, *Reinforcement algorithm* などの方法が実際的でかつ計算機にのりやすい。この点，*Bayes approach* は計算が繁雑である。

参考文献

1. R. Bellman, Adaptive Control Process, A Guided Tour, Princeton Univ. Press, 1961
2. 深尾，適応制御過程の諸性質について，電気試験所彙報，Vol. 28, No. 1, pp. 1-19, 1964
3. 深尾，State, System-Identificationについて，数理解析研究所講究録 8, pp. 39-70, Feb. 1966
4. A. A. Fel'dbaum, Optimal Control System, Academic Press, 1965
5. A. A. Fel'dbaum, Theory of dual control, I, II, III, IV. Automation and Remote Control, Vol. 21, No. 9, pp. 1240-49, No. 11, pp. 1453-64 (1960); Vol. 22, No. 1, pp. 3-16, No. 2, pp. 129-143 (1961)
6. M. Aoki, Optimization of Stochastic Systems, Academic Press, 1967
7. D. Sworder, Optimal Adaptive Control Systems, Academic Press, 1967
8. P. Dorato and R. F. Drenick, Optimality, Insensitivity and Game Theory, L. Radnović (ed), Sensitivity Method in Control Theory, Pergmon Press, pp. 78-109, 1966
9. H. Robbins and S. A. Monro, A stochastic approximation method, Ann. Math. Statist., 22, pp. 400-407, 1951
10. E. Kiefer and T. Wolfowitz, Stochastic estimation of the maximum of a regression function, Ann. Math. Statist., 23, pp. 462-466, 1952

11. A. Dvoretzky, On stochastic approximation, Proc. of the Third Berkeley Symp. on Math. Stat. and Prob., Vol. 1, pp. 39-55, 1956

12. Ya. Z. Tsypkin, Adaptation, Training and Self-Organization in Automatic Control Systems, Automation and Remote Control, Vol. 27, No. 1, pp. 16-51, 1966

13. Ya. Z. Tsypkin, Optimization, Adaptation and Learning in Automatic Systems, Tou(ed), Computer and Information Science II, Academic Press, pp. 15-32, 1967

14. K. S. Fu, Stochastic Automata as Model of Learning Systems, Tou(ed), Computer and Information Science II, Academic Press, pp. 177-191, 1967

15. Y. T. Chien and K. S. Fu, On Bayesian Learning and Stochastic Approximation, IEEE Trans. on System Science and Cybernetics, Vol. SSC-3, No. 1, pp. 28-38, 1967

16. Y. Sawaragi et al., Statistical Decision Theory in Adaptive Control Systems, Academic Press, 1967

17. R. Bush and K. Estes, Studies in Mathematical Learning Theory, Stanford Univ. Press, 1959

18. D. Blackwell and M. A. Girshick, Theory of Games and Statistical Decisions, Wiley, 1954

1:30

Appendix I

[Dvoretzkyの定理]

X_1, Y_n を normed linear space $\mathcal{N}$ の element とする。また θ を $\mathcal{N}$ の element, $T_n(X_1, X_2, \cdots, X_n)$ を $\mathcal{N}\times\mathcal{N}\times\cdots\times\mathcal{N}$ から $\mathcal{N}$ への measurable transformation とする。

(D·1)

$$\| T_n(X_1, X_2, \cdots, X_n) - \theta \| \leq F_n \| X_n - \theta \| \quad (n=1,2,\cdots)$$

(D·2) $\prod_{n=1}^{\infty} F_n = 0$

なる $F_n > 0$ が存在する。

(D·3) $X_{n+1} = T_n(X_1, \cdots, X_n) + Y_n$

とおくとき,

(D·4) $E\{\|X_1\|^2\} + \sum_{n=1}^{\infty} E\{\|Y_n\|^2\} < \infty$

(D·5) すべての measurable transformation $\varphi(X_1, X_2, \cdots X_n)$ に対して $E\{\|\varphi(X_1, X_2, \cdots, X_n) + Y_n\|^2\}$

$$\leq E\{\|\varphi(X_1, X_2, \cdots, X_n)\|^2\} + E\{\|Y_n\|^2\}$$

以上, (D·1)~(D·5)を満足すれば,

$$\lim_{n\to\infty} E\{\|X_n - \theta\|^2\} = 0$$

$$P\left\{\lim_{n\to\infty} \|X_n - \theta\| = 0\right\} = 1$$

が成立する。